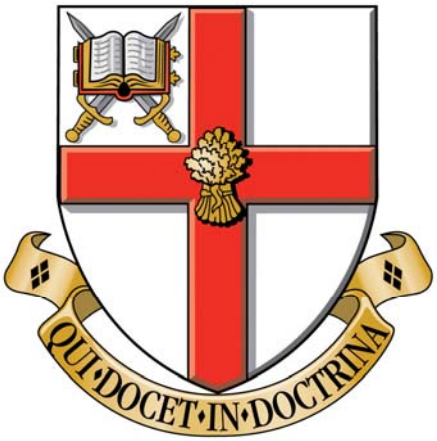

University of Chester

Author(s): Edward Winter; Roger Eston; Kevin L Lamb

Title: Statistical analyses in the physiology of exercise and kinanthropometry

Date: 2001

Originally published in: Journal of Sports Sciences

Example citation: Winter, E., Eston, R., & Lamb, K. L. (2001). Statistical analyses in the physiology of exercise and kinanthropometry. *Journal of Sports Sciences*, 19(10), 761-775

Version of item: Author's post-print

Available at: http://hdl.handle.net/10034/26052

Journal of Sports Sciences

Research Methods Supplement

Online Publication Date: 01 October 2001

http://dx.doi.org/10.1080/026404101317015429

Statistical Analyses in the Physiology of Exercise and Kinanthropometry

E M Winter[1], R G Eston[2] and K L Lamb[3]

[1] The Centre for Sport and Exercise Science, Sheffield Hallam University, Collegiate Hall, Sheffield S10 2BP, [2] School of Sport, Health and Exercise Sciences, University of Wales, Bangor LL57 2 EN and [3] Department of Physical Education and Sports Science, University of Chester, Chester CH1 4BJ

Address for correspondence:

Professor Edward M Winter

The Centre for Sport and Exercise Science

Sheffield Hallam University

Collegiate Hall

SHEFFIELD S10 2BP

Telephone: 0114 225 4333

Fax: 0114 225 4341

e-mail: e.m.winter@shu.ac.uk

Abstract

Research into the physiology of exercise and kinanthropometry is intended to improve our understanding of how the body responds and adapts to exercise and if studies are to be meaningful, they have to be well designed and analysed. Advances in personal computing have made available statistical analyses that were previously the preserve of elaborate main-frame systems and increased opportunities for investigation. However, the ease with which analyses can be peformed can mask underlying philosophical and epistemological shortcomings. The purpose of this review is to examine the use of four techniques that are especially relevant to physiological studies: first, bivariate correlation and linear and non-linear regression; second, multiple regression; third, repeated-measures analysis of variance; and fourth, multilevel modelling. The importance of adhering to underlying statistical assumptions will be emphasised and ways to accommodate violations of these assumptions will be identified.

Key Words: statistics, correlation, regression, repeated-measures designs, multilevel modelling

1. Introduction

The physiology of exercise is the study of how the body responds and adapts to exercise and as with other discipline areas, research is driven by attempts to formulate and answer research questions. Hypotheses are set and then tested by statistical analyses and usually, one of two broad approaches is taken: either possible *differences* between groups or alternatively, possible *relationships* between variables are explored. Whichever approach is taken, two hypotheses are stated: first, the experimental or research hypothesis H_1; and second, the null hypothesis H_0. Consequently, a knowledge and understanding of statistical analyses is a prerequisite for sound research because the investigator can select appropriate techniques to answer the research question or questions that are posed.

Testing tends to follow Popper's falsification principle (Popper, 1963) which states that before a particular hypothesis can be accepted, the opposite has to be rejected. As a result, it is usually H_0 that is tested and according to the outcome, this hypothesis is either accepted or rejected. If H_0 is rejected, H_1 is accepted whereas if H_0 is retained, H_1 is rejected. Acceptance or rejection of H_0 depends on the probability that the outcome could have occurred simply by chance. By convention, the 5% level of probability *i.e.* $P \leq 0{\cdot}05$ tends to be the cut-off level. However, it is important to note that probability levels are not sacrosanct, they are selected by an investigator according to circumstances (Franks and Huck, 1986). The rejection of H_0 when it is actually true is called a Type I error. Conversely, the acceptance of H_0 when it is actually false is called a Type II error. The probability of making a Type I error is symbolised by α whereas the probability of making a Type II error is symbolised by β. If α is made more stringent it becomes harder to reject H_0 so the chances of committing a Type I error are reduced. However, by doing so, the chances of committing a Type II error are now increased. As a result, the importance of calculating the power of a test is highlighted. Power is $1 - \beta$ and is linked to effect size (Thomas and Nelson, 1996; Vincent, 1999) which is a value used to assess practical as opposed to statistical significance of test results.

Measures of effect size incorporate variance of data and typically represent a ratio of the strength of the effect to the variability in the data. Power and effect size are also linked with calculations of sample size which should occur before experiments start (Thomas and Nelson, 1996; Sokal and Rohlf, 1995). These calculations are influenced by α, β, effect size and the variance of proposed groups. For instance, for a given α and effect size, a required β can be achieved with a comparatively small sample if the variance of the group is low. On the other hand, if the variance is likely to be high, a larger sample size would be required. Calculations of sample size are intended to make the selection of subjects as economical as circumstances allow.

The development of personal computing has led to tremendous advances in the ease with which statistical analyses can be performed. Highly complex procedures with their outcomes and actual probability levels are routinely presented. However, it is also important to recognise that the ability to interpret these outcomes is a vital requirement of sound research and this becomes more exacting as statistical packages increase their range of procedures. Statistical analyses in the physiology of exercise are designed to contribute to improving our understanding of how the body responds and adapts to exercise. In particular, they help to identify mechanisms that explain rather than simply describe performance. Investigators already have stern challenges: not only do they have to be competent in physiological and biochemical measuring techniques so as to minimise errors in measurement but they also have to contend with the variability of biological systems. In studies into mechanisms of fatigue, they often have to use demanding exercise protocols that challenge the recruitment and retention of subjects. In such studies where data might be recorded over extensive periods of time with multiple dependent variables, it is difficult to satisfy underlying statistical assumptions and avoid committing Type I or Type II errors. Nevertheless, it is important that investigators are aware of potential pitfalls because they are then better prepared to meet the challenges that are presented.

The application of statistics to the solution of biological problems has been called *biometry* (Sokal and Rohlf, 1995 *p*. 1), a term that is derived from the Greek *bios* ("life") and *metron* ("measure"). It is also sometimes called biological statistics or simply, biostatistics. The origin of modern statistics in general can be traced to the seventeenth century (Sokal and Rohlf, 1995) and biometry grew out of this study. Notable in this field was the Belgian astronomer and mathematician Adolphe Quetelet (1796-1874), after whom the Quetelet (ponderal) index is named, but according to Sokal and Rohlf (1995), the father of biometry and the branch of genetics called eugenics is Francis Galton (1822-1911). It was he who applied statistical methods to the analysis of biological variation, especially through the use of correlation and regression techniques. Other leading figures have been: Florence Nightingale (1820-1910), perhaps better known for her nursing although she was also an excellent mathematician; Karl Pearson (1857-1936), after whom Pearson's product-moment correlation coefficient is named; W.F.R. Weldon (1860-1906) a zoologist and contemporary of Pearson; and Ronald A. Fisher (1890-1962) who has made many contributions to the development of statistical theory.

The purpose of this section is not to review statistical procedures in general because there are sound texts that do this already (Snedecor and Cochran, 1989; Sokal and Rohlf, 1995; Tabachnick and Fidell, 1996). The aim is to highlight biometric techniques that are especially relevant to the physiology of exercise and kinanthropometry. This complements other sections in this supplement such as Selected issues in the design and analysis of sports performance (Atkinson and Nevill, this issue) that addresses issues that are also relevant to physiology *e.g.* reproducibility of measures and limits of agreement (Bland and Altman, 1986); and effect size in Research design and statistics in biomechanics and motor control (Mullineaux *et al*., this issue). This review focusses on four areas: first, bivariate correlation and linear and non-linear regression; second, multiple regression; third, repeated-measures analysis of variance; and fourth, multilevel modelling.

2. Bivariate Correlation and Regression

2.1 **Correlation**

At the outset, it is important to recognise that correlation and regression are often confused. This is unfortunate because they are not synonymous. Reasons for the confusion will become apparent when the principles that underpin regression are examined but the fundamental purpose of correlation is to, ". . . determine whether two variables are interdependent, or **covary** - that is, vary together" (Sokal and Rohlf, 1995 *p*. 557). Because there are two variables, this is called bivariate correlation. No distinction is made between independent and dependent variables and no attempt is made to predict one variable from knowledge of another. Usually, an independent variable is termed *x* and a dependent variable is termed *y.* However, in correlation where no such distinction is made, it has been suggested that the variables should be termed Y_1 and Y_2 (Sokal and Rohlf, 1995).

When data are parametric *i.e.* they are measured on interval or ratio scales, are continuous and include a bivariate normal distribution, the degree of association between variables is usually calculated as Pearson's product-moment correlation coefficient, *r*. This coefficient ranges from 0 to 1 for positive relationships in which as Y_1 increases so too does Y_2, and 0 to -1 for relationships in which one of the variables decreases as the other increases. The value r^2 x 100 is the coefficient of determination and expresses the variance in one variable that can be attributed to its relationship with the second variable. The value (1 - r^2) x 100 is sometimes known as the coefficient of non-determination and the square root of this coefficient *i.e.* $\sqrt{1-r^2}$ x 100, is known as the coefficient of alienation and is a measure of the lack of association between variables Y_1 and Y_2. When data are non-parametric *i.e.* they do not have a normal distribution, tests of association can be made using Spearman's rank order correlation coefficient, ρ, or Kendall's τ (Sokal and Rohlf, 1995).

Exploration of relationships between physiological and performance measures begins to identify possible mechanisms that could explain rather than simply describe performance. Clearly,

this increases the precision with which the physiological status of individuals can be assessed. This is especially applicable to sport and exercise science. However, the value of *r* in particular is influenced by the ranges of Y_1 and Y_2 (Smith, 1984). Large ranges in one or both variables can produce high values of *r* whereas low ranges can depress *r.* This is illustrated by Sale (1991) and shows how the precision of the relationship between Y_1 and Y_2 can be misrepresented. This leads to ways in which such precision can be assessed and how one variable can be predicted from knowledge of another *i.e.* regression.

2.2 **Regression**

Regression involves the identification of functional relationships between variables. "Functional" is in the context of relationships that are mathematical as opposed to biological but might and indeed probably would lead to suggestions of cause-and-effect. However, precise demonstrations of causality have to be achieved through established procedures of scientific method (Sokal and Rohlf, 1995) and not simply by what could be chance associations between variables. When the relationship between two variables is examined, it is called bivariate regression.

The literal meaning of regression is returning or going back. However, in the context of a bivariate relationship a line of best-fit is constructed that is based on the principle of least-squares that describes pairs of data points. Individual data points lie away from the regression line and the line is the one that minimises the distance of these points. Because some points are above the line and some below, the difference between each actual and fitted value is squared, hence the term least-squares. In this and in most cases, it is the least-squares in a vertical direction that are minimised. This *Y* on *X* regression is called Model I regression (Sokal and Rohlf, 1995) but it is important to recognise that such minimisation is not restricted to the vertical and other models will be considered shortly.

By convention the regression line is expressed in the form $Y = a + bX$ where a is the intercept *i.e.* the point where the line crosses the ordinate and b, the regression coefficient, is the gradient of the line. Concerted applications of scaling *i.e.* ways to adjust physiological and performance variables for differences in body size (D'Arcy-Thompson, 1917; Tanner, 1949; Huxley, 1932; Packard and Boardman, 1987) to sport and exercise science (Katch, 1973; Nevill and Holder, 1995; Rogers *et al*., 1995; Winter, 1996) have highlighted the need for a sound knowledge and understanding of principles that underpin linear and non-linear regression models (Batterham and George, 1997).

Model I regression is based on four assumptions (Sokal and Rohlf, 1995): first, that the independent variable X is measured without error *i.e.* each value of X is "fixed", whereas the dependent variable Y is a random variable; second, that the relationship between Y and X is linear and is described by the equation $Y = a + bX$; third, for any given value of X the Y's are independently and normally distributed; and fourth, error about the regression line is homoscedastic *i.e.* the variance about the regression line is constant and does not depend on the magnitude of X or Y.

Where one or more of these requirements cannot be met, what regression model should be used? Sokal and Rohlf (1995) refer to alternatives as Model II regression but the question has been the subject of considerable debate (Ricker, 1973; Kuhry and Marcus, 1977; Seim and Sæther, 1983; Rayner, 1985) so what follows is a brief outline of the main points. Figure 1 is a simple example of the principles.

* * * * Figure 1 near here, please * * * *

The data are unpublished test-retest results of an anthropometric assessment of lean thigh cross-sectional area in women. Pearson's product moment correlation coefficient was 0.869 (P =

0.001). Regression line A is the traditional Model I or Y on X regression that minimises the sum-of-squares in a vertical direction. The equation for this line is:

$Y = 34.5 + 0.777X$ Standard error of the estimate = 7.7 cm^2

Coefficient of variation = 5.0%

The standard error of the estimate is the square root of the variance about the regression line except that the denominator is $n - 2$ rather than $n - 1$. The coefficient of variation about regression is the standard error of the estimate expressed as a percentage of the mean of Y and can be a useful measure of the variability about regression. Line B represents X on Y in which the sum-of-squares in a horizontal direction is minimised. The equation for this line is:

$$Y = -4.6 + 1.029X$$

and is obtained simply by transposing the variables X and Y when the equation for the regression line is determined. Note that the slope of this line is greater than the one for Y on X. Line C represents what has been termed by Ricker (1973) the geometric mean regression. Ricker (1973) also explains how to calculate the equation for this line. The line bisects A and B and applies when neither variable can properly be considered to be dependent or independent and when one variable is not fixed; both are subject to error. The equation for C is:

$$Y = 30.5 + 0.969X$$

As the relationship between the variables becomes less pronounced, the difference between the slopes of the three regression lines just outlined will increase.

Line C has also been called the reduced major axis (Kermack and Haldane, 1950) and la relation d'allométrie (Teissier, 1948). It is rare to see this regression line in sport and exercise science but Winter *et al*. (1996) used it in their study of optimised and corrected peak power output during cycle ergometry. Another line that has been considered in the context of linear regression is the principal axis (Snedecor and Cochran, 1989). This is the axis of the correlation elipse that represents the bivariate normal distribution but does not feature as an actual regression line.

The importance of being aware of different linear regression models is accentuated when non-linear relationships are investigated. At first sight this seems curious but reasons soon become apparent. Non-linear relationships take different forms. They can be polynomial in which the independent variable can be raised to an integer power such as x^2, x^3, x^4 and so on. Other forms of exponential relationships are exemplified by oxygen uptake responses to the onset of exercise (Whipp, 1996) and the relationships between various physiological and peformance measures and body size that are allometric (Schmidt-Nielsen, 1984).

Scaling involves the investigation of possible allometric relationships that are of the general form:

$Y = a \times X^b \times \varepsilon$ where a = constant multiplier

b = the exponent

ε = error term

The error term, ε, is multiplicative *i.e.* as X increases so too does ε. The allometric model can be linearised by transforming X and Y to natural logarithms so producing:

$$\ln Y = \ln a + b \ln X + \ln \varepsilon$$

Identification of b can be used to construct power function ratios *i.e.* Y/X^b (Winter and Nevill, 1996) which allow inter group comparisons to be made with measures adjusted for differences in body size. Clearly, the numerical value of b depends on the regression model that is used to construct the regression line and this is probably the major reason why extensive debate has taken place about which is the appropriate model to use (Ricker, 1973; Rayner, 1985; Sokal and Rohlf, 1995).

However, indiscriminate transformations should not be made because they can be troublesome (Batterham and George, 1997). For instance, transformation of unskewed raw data could introduce skewness and might mask a non-linear log-log relationship. Most statistics packages can perform appropriate checks *e.g.* by producing normal probability plots to examine the distribution of each variable to see if violations have occurred. Similarly, the distributions of residuals *i.e.* the difference between actual and predicted values can also be examined. Interest in scaling in sport and exercise science has increased markedly in the last decade and built on well established procedures elsewhere in biological science (Schmidt-Nielsen, 1984). Studies and reviews have addressed theoretical issues (Nevill, 1994; Nevill and Holder, 1995; Batterham and George, 1997), maximum oxygen uptake and body size (Nevill and Holder, 1994), comparisons between the performance capabilities of men and women (Winter *et al.*, 1991; Vanderburgh *et al*., 1995; Eston *et al*., 1997); and children and adults (Eston *et al*., 1993; Winter, 1996; Welsman *et al*., 1996; Rogers *et al*., 1995).

Analyses of allometric relationships have become increasingly complex and look set to continue in sport and exercise science. It has also become clear that they should correctly specify and confirm underlying model assumptions (Batterham and George, 1997; Nevill and Holder, 1999) to ensure physiological mechanisms can be investigated meaningfully.

3. Multiple Regression Analysis

The bivariate regression concept is frequently expanded to situations where the dependent or criterion variable (*Y*) is related to two or more independent (predictor) variables. The independent variables are usually designated by the symbol x_1, x_2, x_3, etc., to calculate a predicted score (*Y'*) on *Y*, using the formulae: $Y' = a + b_1x_1 + b_2x_2 \ldots b_kx_k$, where b_1, b_2, …..b_k are regression coefficients or standardised units called beta weights that provide a weight to the predictor variable according to the relative contributions to the prediction of Y.

The coefficient of multiple correlation (*R)* is an index of the accuracy of the equation. It can be thought of as a simple Pearson correlation between the actual *Y* scores and the predicted *Y* scores if the multiple regression equation was used to obtain a score for each participant in the original group from which the prediction equation was derived (Huck *et al*., 1974). The coefficient of determination (R^2), like the bivariate equivalent (r^2), is frequently reported to calculate the proportion of *Y* variance that is predictable on the basis of scores on the predictor variables. For example, if *R* = 0.80, 64% of the *Y* variance can be predicted from the set of *X* scores. It should be noted that R^2 is a positively biased estimate. The adjusted R^2 which corrects this bias and therefore has a lower value, is given by default in SPSS for example. (Kinnear and Gray, 1997).

The purpose of multiple regression is to find the most satisfactory solution to the prediction of *Y*. This is the solution that produces the lowest standard error of the estimate (SE_E) (Vincent, 1999). This can be interpreted as the standard deviation of all the errors, or residuals, when predicting *Y* from *X* and can be calculated from the standard deviation of *Y* (SD_Y and *R*, using the formula: $SE_E = SD_Y (1-R^2)^{0.5}$. The SE_E is reported in most studies and provides a means by which confidence intervals for the accuracy of the prediction can be calculated. That is, 68% of all the

errors in the prediction will be between ±1 x SE_E, 95% will be between ±1.96 x SE_E and 99% will be between ±2.58 x SE_E. A simple application of this concept is demonstrated by Vincent (1999).

Multiple regression analysis is commonly used in exercise physiology and kinanthropometry research. For example, studies to determine the best combination of skinfold sites (Jackson and Pollock, 1978; Eston *et al*., 1995) or bioelectrical impedance values (Houtkooper *et al*., 1989; Eston *et al*., 1993) for predicting body composition, or the best combination of variables for predicting energy expenditure in children (Eston *et al*., 1998) provide typical examples. The following example is taken from a recent study to determine which combination of measurements recorded by various activity monitors (pedometry, single and three dimensional accelerometry and heart rate telemetry) provided the best prediction of total activity (walking, running, hopping, catching and jumping) as measured by oxygen uptake ($\dot{V}O_2$) in 40 boys and girls.

Raw score formula:

$\dot{V}O_2$ = 6.21 + 0.01 (Tritrac 3D score) + 0.176 (HR)

R= 0.92, R^2 = 0.85,

SE_E = 9.7 ($ml.kg^{-0.75}.min^{-1}$) or 17% of the mean.

Standardized formula:

$\dot{V}O_2$ = 6.21 + 0.735 (Tritrac 3D score) + 0.225 (HR)

In this study, multiple regression analysis revealed that the combination of three dimensional accelerometry (Tritrac 3D score) and heart rate (HR) was the strongest predictor of energy expenditure. The R^2 indicates that in combination they accounted for 85% of the variance in energy expenditure. The second equation was also of interest in this study. Regression coefficients cannot be compared against one another because the predictor variables in the prediction equation are not

on the same scale of measurement. For example, the accelerometer accumulated movement counts in three dimensions, whereas heart rate was measured in beats per minute. As the researchers were interested in knowing which of the independent variables was the best predictor, the regression coefficients were converted into comparable standardised units called 'beta weights' calculated from standardised scores (*Z*). This procedure is computed automatically in many statistics programs such as SPSS. The predictor variable that has the largest beta weight, regardless of whether the beta weight is positive or negative, is the best predictor of *Y*. The beta weights in the standardized formula indicate that three dimensional accelerometry was a better predictor of energy expenditure in comparison to heart rate telemetry.

Venn diagrams can be used to explain the concepts of 'shared' and 'unique' variance and how the influence of independent variables in multiple regression is determined (Tabachnick and Fidell, 1996; Vincent, 1999). Figure 2 presents a hypothetical example to show how the amount of variance in percentage body fat (as measured with a valid criterion technique) can be explained by the variance in other independent variables. In this example, the final multiple regression equation would include the sum of skinfolds because this accounts for the greatest proportion of variance in percent fat (about 65%). It would not include height and body mass because they do not account for much variance in percent fat and more importantly, the amount of variance accounted for by these factors is shared by the variance in the sum of skinfolds and the waist to hip ratio. The latter factor accounts for about 25% of the variance in percent fat, of which about 15% is *unique* variance. That is to say, the waist to hip ratio accounts for an additional 15% of the variance in percent fat which is not accounted for by the sum of skinfolds. In addition, although self-reported physical activity contributes to the variance in percent fat, most of this is already accounted for by the sum of skinfolds. The proportion of unique variance it shares with percent fat is negligible and therefore would not be included in the final multiple regression equation.

The relationship between predictors has implications for the choice of measures included in a study. Ideally, the relationship between predictor variables should be low, so that each factor explains a unique variance which is not common to the other predictor variables. However, this situation is not common in practical research. Thus, to produce the equation containing the best subset of predictors and which has the highest predictive value and the lowest SE_E, investigators should select independent variables that are highly related to the criterion measure and not related to each other.

There are a number of methods for deriving a subset of predictor variables which include stepwise and hierarchical procedures (Munro, 2001). The most common form of multiple regression analysis reported in physiology and kinanthropometry research involves forward stepwise regression procedures. In this process the computer establishes which factor is the best predictor and lists the bivariate solution first. The equation is then expanded in a step-by-step procedure as each independent variable is added to the equation. The selection criterion for entry into the regression equation is frequently set at 0.05. In this case, a variable has to account for a significant (P<0.05) amount of variance to be included in the analysis. The selection ends when there is no further significant contribution by the remaining independent variables. With this procedure, once a variable is in the equation, it is not removed. No attempt is made to reassess the contribution of a variable once other variables have been added.

An alternative method which is superior to the forward solution, is backward stepwise regression, although this is a lesser used procedure in physiology research. The procedure starts with the overall R^2 which is generated by putting all of the independent factors into the equation. Each variable is then deleted one at a time to determine if the R^2 drops significantly. As Munro (2001) has outlined, each variable is tested to see what would happen if it were the last one entered into the equation. She illustrated the procedure with four independent variables to show how the

differences in R^2 would be listed:

$R^2_Y.1234 - R^2_Y.234$ Tests for variable 1

$R^2_Y.1234 - R^2_Y.134$ Tests for variable 2

$R^2_Y.1234 - R^2_Y.124$ Tests for variable 3

$R^2_Y.1234 - R^2_Y.123$ Tests for variable 4

If there is a significant drop in R^2, that variable contributes significantly and is not removed. However, if there is not a significant drop in R^2 the variable is removed. In this case, each of the three remaining variables is then tested to see if it would contribute significantly if entered last. The analysis continues until all variables in the equation contribute significantly if entered last. In this way, unlike the forward solution, the backward solution reassesses the contribution of a variable once other variables have been added and is therefore considered to be the best of the stepwise regression methods (Nevill and Atkinson, 2001). Using the data from the study of Cockerill *et al*. (1991), on factors which influenced performance time in cross-country runners, they illustrated how the choice of backward or forward regression procedures may affect the number of predictor variables in the final solution.

An important consideration in these forms of multiple regression analysis is the danger of including variables that have no sound theoretical relationship with the criterion, and lead to 'fishing trips' with the data. Because the order of entry is based on statistical rather than theoretical rationale, the techniques are criticised for capitalising on chance (Munro, 2001). An alternative and theoretically stronger approach is hierarchical regression analysis (Munro, 2001) in which the predictor variables are entered by the investigator in a specified order, determined *a priori*, according to the underlying theory. The procedure therefore requires a sound grasp of the underlying theory as already acknowledged by Cohen and Cohen (1983) and Biddle *et al*. (this issue). This implies that

the investigator must understand the rationale and justification for including the independent variables.

Hierarchical regression analysis, also known as sequential regression (Tabachnick and Fidell, 1996), is used much less frequently in sport and exercise physiology research. However, the method was applied recently in studies to determine the relationship between energy expenditure, movement and body fat in children (Eston *et al*., 1998; Rowlands *et al*., 1999; Louie *et al*., 1999). In these studies, the procedure was most useful to determine the additional proportion of variance of *Y* which could be explained by the deliberate inclusion or omission of variables in the analysis. For example, in determining which method of activity analysis was the best predictor of energy expenditure in children, Eston *et al*. (1998) alternated the order of entry to illustrate the amount of variance accounted for by the second measure over and above that already accounted for by the first measure. This analysis revealed the regression equations listed above. Heart rate added 2% ($P<0.01$) of the variance to that already explained by the 3D accelerometer, whereas the 3D accelerometer was responsible for an extra 21% ($P<0.01$) in addition to that already accounted for by heart rate.

It is highly likely that the regression equation will be less accurate when used with new people, owing to differences between the participants and the original sample used to derive the equation. Investigators can however use a procedure known as 'cross-validation' to determine if their equations have a chance of being successful when applied to the new group of individuals. Huck *et al*. (1974) list four stages to this procedure: 1. The original sample is split into two subgroups; 2. One of the sub groups is used to develop a prediction equation; 3. The equation is used to predict a criterion score for each person in the second subgroup; and 4. The predicted score is then correlated with the actual criterion score (see Eston *et al*., 1995, for a simple example of this type of procedure).

An important consideration when using multiple regression procedures, and which has been addressed previously in this Journal (Bartlett, 1997), concerns the ratio of the sample size (n) to the number of predictor variables (p). Some studies use far too few participants and too many predictor variables. As the expected value of R for random data can be estimated from $p/(n\text{-}1)$, the ratio of participants to predictor variables should be no less than 5:1 and ideally about 20:1 (Tabachnick and Fidell, 1996; Vincent, 1999). For stepwise regression, an even greater ratio of 40:1 is recommended (Vincent, 1999).

Vincent (1999) lists other cautions and assumptions which must be considered by the researcher. Outliers, cases with excessively large residual values, which may be extreme cases in the population or the result of error, should be found and corrected, because they will bias the prediction equation. An assessment of the scattergram between variables is helpful to identify outliers. An important tool for checking the assumptions is residual analysis. A residual is the difference between the actual score and the predicted score. In other words, if the analysis was perfect, the sum of the squares of residuals would be zero. Residuals of each value of X should be checked to ensure that the data are normally distributed around the best fit line. Normal distribution of the residuals can be checked from a histogram of the standardised residuals. A normal curve is interposed on the standardised residuals. A further plot is a cumulative or probability plot of the standardised residuals. Ideally, points should lie along a diagonal line. If they do not, the residuals are not normally distributed and again, it might be necessary to apply a transformation to the data (Howell, 1997; Kinnear and Gray, 1997). Finally, examination of a scatterplot of the predicted values against the residuals should be used to confirm if the assumptions of linearity and homogeneity of variance have been met. If the scatterplot is crescent- or funnel-shaped, this indicates heteroscedasticity and the data should be screened further or the analysis abandoned (Kinnear and Gray, 1997).

4. Repeated Measures Designs

One of the most prevalent types of research design to be found in the physiology of exercise and sport literature is the within-subjects or repeated measures design in which one or more dependent variables are measured two or more times on one or more groups of subjects. Typically, such research is experimental in nature and concerned with observing whether a specified intervention programme (treatment) has a meaningful effect on a defined measure of performance or physiological function. The popularity of this type of research is due to its economy and statistical power. Compared with an independent groups design, it requires relatively few subjects and is more likely to identify a true treatment effect on a dependent variable due to a lesser amount of between-subjects variability. An example from a recent edition of the *Journal of Sports Sciences* is the study by Peyrebrune *et al*. (1998) that examined the effects of a five-day supplementation of creatine on 50-yard sprint times among elite male swimmers.

From a data analysis perspective, it is usually the mean value of the dependent variable post-treatment that is compared to the mean value pre-treatment, with the null hypothesis (that there is no difference) being tested with an appropriate statistical test. The *repeated measures analysis of variance* is the appropriate test when there are at least two mean values (scores) belonging to one or more groups being compared. Whether the repeated measures analysis of variance is prefixed with the words *one-way*, *two-way*, *three-way* and so on, depends upon the number of independent variables (factors) in the analysis. For example, when a single group is measured twice (pre/post) and under two treatment conditions (treatment/control or placebo), the analysis is *two-way* and there are four means for comparison. Thomas *et al.* (1998) employed a two-way repeated measures analysis of variance in their study of the effects of wearing a mouthpiece and/or nasal strip on measures of anaerobic performance. A mixed group of adults (n = 15) completed the Wingate anaerobic test on six randomly ordered occasions (providing six means for comparison) whilst

wearing (or not) nasal strips and mouthpieces. Their performance measures (dependent variables) in this 2 (mouthpiece/no mouthpiece) x 3 (no nasal strip/placebo nasal strip/nasal strip) *fully repeated measures* design (since each subject was exposed to all conditions), were peak anaerobic power and anaerobic capacity.

Likewise, when two separate (independent) groups are formed and are each measured on two or more occasions, there are at least four mean scores for comparison. This *mixed factorial* or *mixed model* design, combines a within-subjects factor (time) and a between-subjects factor (group). It is distinctive in that it requires subjects to be measured on fewer occasions than a fully repeated measures design. An example is a study into the effect of a 4-month physical training programme on measures of total and visceral fat in obese children (Owens *et al.*, 1999). All subjects (n = 74) were measured at the start and at the end of the study, but only half of them (the experimental group) were given the physical training. As above, the interaction term in the repeated measures analysis of variance (RM ANOVA) was used to assess whether the fat values at the end of the study were significantly different (lower) in the experimental group compared to the untrained control group. A further example of a mixed model design is the study by Eston and Peters (1999) that compared the effects of cryotherapy on symptoms of exercise-induced muscle damage (EIMD). Symptoms of EIMD were compared in two groups who had performed eccentric muscle actions over a period of three days.

Repeated measures designs in which there is just one independent variable with three or more levels (trials or conditions) are common in physiology research and also lend themselves to RM ANOVA. Whilst non-experimental in nature (being descriptive or quasi-experimental), such designs still allow differences in the dependent variable across the levels of the independent variable to be tested for significance. For example, Norris and Petersen (1998) examined whether oxygen uptake kinetics of elite cyclists changed following endurance training. Mean values of variables such as

maximal oxygen uptake and oxygen uptake at ventilatory threshold measured pre-training, mid-training (after 4 weeks) and post-training (after 8 weeks) were compared using RM ANOVAs. Similarly, a study that addressed the reliability of a protocol for testing endurance performance in runners and cyclists used separate RM ANOVAs to compare several dependent variables measured during four repeated trials (Doyle and Martinez, 1998).

Designs in which there are more than two independent variables and/or the independent variables have more than two levels (conditions or treatments or trials) are less common, but nevertheless evident in physiology research. Two examples are provided by the research of Trappe *et al.* (1995) and Eston *et al*. (1992). In the first (three independent variables and hence a three-way RM ANOVA), thermal responses (core and trunk temperatures) of competitive swimmers to swimming in three water temperatures whilst wearing either a wet suit or swimming suit were monitored every ten minutes for 30 minutes. In the second (two independent variables and hence a two-way RM ANOVA), the heart rate, blood pressure, blood lactate and perceived exertion responses of obese, sedentary women to three levels of exercise intensity were monitored at four stages of a 7-week dietary intervention programme.

This introduction has simply identified the general situations (research designs) in which physiologists are likely to utilise RM ANOVA. In recognising that many texts on research methodology and quantitative statistics give RM ANOVA considerable attention (e.g. Huck and Cormier, 1996; Howell, 1997), what follows is an attempt to summarise its key elements.

4.1 **Important Assumptions for Repeated Measures ANOVA**

As RM ANOVA is a *parametric* test, the dependent variables on which it is used should be measured at the interval or ratio level. The independent variables (factors) should also be categorical. In addition, the fundamental assumptions that relate to the use of ANOVA for fully

between-subjects (independent groups) designs are also relevant to RM ANOVA. This means that, (i) the dependent variable should be normally distributed in the population from which the sample of subjects is drawn, (ii) the variances of the scores obtained in each group or condition (or trial) should be similar, and (iii), the scores of subjects in different groups are independent of each other. Now, whilst it seems that statisticians agree that the between-subjects ANOVA can cope with modest violations to these assumptions, especially if the samples are similar in size and not low in number (> 10), an extra assumption that is specific to repeated measures designs, compound symmetry or *sphericity*, certainly has to be met by the data. Sphericity means that there is homogeneity of covariance *i.e.* correlations among all combinations of trials are equal (Vincent, 1999). If investigators do not check for, and if necessary deal with this assumption, some statisticians hold the view that any conclusions they draw should be disregarded (Huck and Cormier, 1996, *p*. 432).

4.2 The Components of RM ANOVA

As with simple (one-way) independent groups ANOVA, one-way RM ANOVA allows the single null hypothesis among three or more means to be tested whilst keeping the α at the specified level (usually 0.05). However, with RM ANOVA the means are calculated from repeated measurements of the same group of subjects, rather than single measurements of different groups. The consequence of this is manifest in the within-subjects (or error) component of the analysis of variance, that is, as inter-individual variability has been eliminated (by not having different groups of subjects), the denominator of the *F*-ratio is lower than it would otherwise be, increasing the *F*-value. RM ANOVA is therefore a more powerful test than its independent groups counter-part.

If the research design is factorial and has two independent variables (RM factors), a two-way RM ANOVA now allows three hypotheses to be tested in relation to the dependent variable; two relating to the *main effects* of each independent variable, and one relating to the combined or *interaction effect* of the independent variables. In the Thomas *et al.* (1998) study mentioned above,

the RM ANOVA allowed an assessment of whether anaerobic power and anaerobic capacity scores differed according to 1) each RM factor (mouthpiece/no mouthpiece and nasal strip/placebo/no nasal strip), and 2) the combination of the two RM factors. Indeed, it was this interaction that interested the researchers the most as they had hypothesised that the nasal strips would prove to be beneficial when the oral protective devices were being worn. Hence, three *F*-ratios and accompanying tests of significance are generated. Similarly, if the design is mixed factorial (with one RM factor and one independent groups factor), main effects and interaction effects are also tested, but one difference being that RM ANOVA procedures tend to separate out the results under *within-subjects* and *between-subjects* headings. Any interaction effects involving both within- and between-subjects factors tend to appear under the within-subjects heading.

4.3 Checking for Sphericity

As indicated above, a key diagnostic test of the data from RM ANOVA is that of sphericity. If this assumption is violated, the *F*-ratios calculated will be inflated, and the likelihood of a Type I error occurring is enhanced. Sphericity requires that the variances of the repeated measures should be equivalent (homogeneous) and that the bivariate correlations between each repeated measure should also be equivalent (homogeneity of covariance). Whilst these two components could be checked separately, the *Mauchley test* is available in many statistical software packages to give an overall, single assessment of sphericity. If the result from this test is not significant ($P > .05$), the condition can be considered to have been met (termed "Sphericity Assumed" in SPSS for Windows) and the *F*-ratios generated by the RM ANOVA can be accepted.

If the Mauchley test yields a significant result, then the sphericity condition has been violated and the researcher should respond to this. Such a response requires the use of a correction procedure, for example the Greenhouse-Geisser or Huynh-Feldt procedures, that will ultimately make an adjustment to the degrees of freedom (df) which consequently raises the critical (table) value of *F*

and counters the inflated Type I error risk. Which of the correction procedures one should use is debatable and the reader is referred to Howell (1997, pp. 464-466) for a discussion on the features of the main procedures. For more prescriptive guidance, however, the Vincent (1999) text is recommended (*pp*. 178-179). Whatever the choice, Howell (1997, *p*. 466) argues that a correction procedure should *always* be employed, even if the test of sphericity is not significant.

Again, many statistical software packages provide these computations and allow the researcher to report the corrected RM ANOVA results. However, it is notable that many published studies in physiology involving RM ANOVA do not make mention of the sphericity assumption (whether it had been checked and what the consequence was) which could mean that an unknown number of null hypotheses have been falsely rejected due to incorrectly interpreted results.

An alternative approach to dealing with the sphericity assumption is to analyse the repeated measures data with a multivariate technique, in which the repeated measures are not dealt with as levels of a single factor (e.g. trials), but are designated as multiple dependent variables (see Vincent, 1999, *p*.179). Sphericity is not an issue for multivariate analysis of variance (MANOVA), and therefore does not threaten the risk of a Type I error. However, MANOVA requires larger sample sizes and is less powerful than univariate analysis. Moreover, examples of published physiology research that have chosen (and reported) this approach are scarce.

4.4 Interpreting RM ANOVA Output

The capacious statistical software programmes now available to researchers, such as the latest versions of SPSS, Minitab and SAS, allow the data from RM designs to be swiftly analysed and interpreted. Whilst the supporting 'Help' and graphical facilities within such programmes are increasingly more informative, researchers are advised to scour the research methods and statistics texts for examples of computer-generated RM output and how to interpret them. A good example is

the text by Munro (2001) which both presents and explains SPSS (for Windows) output for many of its statistical examples. The following example applies the same logic specifically to data analysed with RM ANOVA in a recently published investigation into the reliability of ratings of perceived exertion (RPE) during progressive treadmill exercise (Lamb *et al*., 1999). The research was non-experimental (having no intervention) and concerned the variability of RPE ratings amongst one group of 16 male athletes over two identical trials and across four exercise levels. Accordingly, the statistical test employed was a two-way (Trials by Levels) RM ANOVA.

Table 1 provides the check on whether the assumption of sphericity has been met (SPSS for Windows). The assumption applies to within-subjects effects that involve more than two repeated measures; hence the value of Mauchley's W (1.00) for Trials (test/retest) is optimal and not tested for significance. For Levels, sphericity has been violated (W = 0.183, $P < 0.0005$), and to a lesser extent for the Trials x Levels interaction effect (W = 0.385, $P = 0.023$). Adjustments to offset these violations are therefore required and Table 2 conveniently presents the results provided by the two popular methods. Each method calculates an *epsilon* correction factor, which is used to modify (reduce) the Sphericity Assumed *d.f.* The smaller this epsilon value, the greater the degree of sphericity violation (Table 1). It is argued (Vincent, 1999, *p*. 179) that the Greenhouse-Geisser adjustment should be used initially as this is the more stringent (conservative) of the correction methods. If the *F*-ratio is significant after this, then the null hypothesis can be rejected. If not, the more lenient Huynh-Feldt adjustment should be evaluated. An alternative recommendation (see Howell, 1997, *pp*. 465-466) is that the Greenhouse-Geisser correction should be used when the average of the two epsilons (an estimate of the 'true' epsilon) is < .75, and the Huynh-Feldt correction used when this average is ≥ .75.

For the present data, the Levels effect is so large that even after the Greenhouse-Geisser adjustment (resulting in a decrease in the df from 3 to 1.514), the *F*-ratio is highly significant ($P <$

0.0005). Similarly, the effect of the Trials x Levels interaction is large enough to withstand the Greenhouse-Geisser correction (*d.f.* reduced to 1.888 from 3) and remain significant (*P* = 0.009).

* * * * Tables 1 and 2 near here, please. * * * *

However, notice that the *P*-value for the uncorrected analysis (Sphericity Assumed) is lower (0.002). Had the interaction effect been less, one can easily imagine how not employing a sphericity correction might lead to the incorrect rejection of the null hypothesis (as the *P*-value approaches 0.05). A further example of where this procedure has been applied is in the study by Eston and Peters (1999).

4.5 *Post-hoc* analyses

For all one-way and factorial RM designs, significant *F*-ratios need to be explored with *post-hoc* tests to identify which pairs of means are significantly different. (The alternative approach to comparing means involves the researcher identifying specific comparisons *a priori* – before the data are collected – and employing tests that are different to those used after the data has been scrutinised. However, as such analyses are rarely adopted in sport and exercise physiology, we will not dwell here on their detail). There are many *post-hoc* tests available (most of which statistical software packages are now capable of applying), and the reader is encouraged again to consult Howell (1997, *pp*. 348-399) for an important discussion of the merits of the different tests. A prominent consideration in Howell's chapter is the increased risk of committing a Type I error due to the researcher making multiple comparisons of group (or trial) means, and how well individual *post-hoc* test control this so-called *familywise error rate* (FER). In addition, Howell highlights the simultaneous loss of power (increased risk of a Type II error) that occurs with certain *post-hoc* tests as they control the FER.

One of the most popular *post-hoc* procedures is Tukey's Honestly Significant Difference (HSD) test[1]. For example, Morris *et al.* (1998) used this form of *post-hoc* test to establish how physiological and metabolic measures varied over time in response to intermittent shuttle running in two environmental conditions. Likewise, El-Sayed *et al.* (1997) used the Tukey test to isolate the effects of carbohydrate ingestion from a placebo on endurance cycling performance over specified time periods.

Now, whilst Howell (1997) supports the use of the Tukey test, especially if the researcher has a large number of means and may wish to make many pairwise comparisons, he advocates the Newman-Keuls procedure in situations where there are fewer than five comparisons on account of it being a more powerful test than the Tukey. However, as the Newman-Keuls has less control over the FER than the Tukey, Howell's preferred *post-hoc* procedure is the Ryan procedure (also identified in statistical software by the acronym REGWQ) which is seen as a compromise test in terms of power and FER control. The well-known Scheffé test is not recommended for pairwise comparisons on account of the high critical values it requires for significance (*i.e.* it is too conservative).

As the Tukey test is so widely used, it is worth noting the concern expressed by Stevens (1996, *p*. 463) that its use is only justified in RM designs when sphericity has been met. This is because the Tukey test does not control the FER too well when the sphericity assumption has been violated, and the likelihood of a Type I error is enhanced. Stevens (1996, *p*. 466) and Howell (1997, *pp*. 471-2) advocate the use of multiple dependent *t*-tests with Bonferroni adjustments (nominated *p*-value divided by the number of *t*-tests). Even with severe violations of sphericity (epsilon = 0.50) this approach was shown to keep the FER at the nominated level. Accordingly, the significant Trials x

[1] Unfortunately, the Tukey and other *post-hoc* tests that use a pooled error term for RM ANOVAs is not available in SPSS. Such procedures must be calculated by hand, or replaced by another *post-hoc* approach, such as that of using dependent *t*-tests with a Bonferroni correction.

Levels interaction referred to in Table 1 was followed-up with Bonferroni-adjusted dependent *t*-tests. These adjustments are intended to reduce the chances of a Type I error occurring but in so doing, the likelihood of a Type II error is increased. For this reason, the use of Bonferroni adjustments has been questioned (Perneger, 1998).

Another *post-hoc* analytical technique is trend analysis (Tabachnick and Fidell (1996) Although this is rarely used in physiology of exercise and kinanthropometry research, it is nevertheless worthy of mention. The principle is straightforward: it is possible for comparisons of means over time to suggest no difference between say, two groups. However, each comparison is taken in isolation yet it could well be that one group was getting slightly better while another was slightly deteriorating. Changes in each group could be subtle, but consistent, and might not be detected by simple comparisons. It is in this situation that analysis of the trends would be appropriate. This also illustrates the value of simple data plots so that the researcher can visually inspect data and so identify, perhaps albeit subjectively, particular trends but who can then move on to specific hypothesis-driven analyses.

4.6 Non-parametric RM ANOVA

If the data from a single factor study do not warrant the parametric repeated measures analysis of variance - due to small or uneven sample sizes and/or the fundamental assumptions of ANOVA being severely violated - the Friedman ANOVA test can be applied as a non-parametric equivalent of the RM ANOVA. The Friedman ANOVA test converts scores into ranks and computes the sum of ranks for each repeated measurement. The null hypothesis that the sums of ranks assigned to each measurement are not significantly different is then tested. If the derived statistic (chi-squared) is large enough to allow the rejection of the null hypothesis, then *post-hoc* analysis can be performed to make multiple pairwise comparisons. Here the researcher can adopt one of two approaches; 1) calculate the critical difference (for significance) by using formulae such as those

presented by Siegel and Castellan (1988, *pp*. 180-181), or 2) employ another non-parametric test, the Wilcoxon matched pairs test, for pairwise comparisons of the repeated measures. From a data analysis perspective, this second approach is easier since it can be achieved within most statistical software programs.

Although rare in the physiology literature, an example of the use of the Wilcoxon matched pairs test as a *post-hoc* procedure is provided by Gill *et al.* (1998). This study examined the effects of type of exercise on postprandial lipaemia. One part of the analysis concerned whether fasting glucose, insulin and nonesterified fatty acid concentrations differed between conditions of no exercise, continuous exercise and intermittent exercise. As the sample size was small (n = 6), the Friedman ANOVA and the Wilcoxon matched pairs tests were appropriately employed. However, it seems that the researchers overlooked the fact that when multiple Wilcoxon tests are performed, the Bonferroni adjustment to offset the inflated risk of a Type I error should also be employed.

5. Multilevel modelling

Multilevel modelling is a form of multiple regression that can analyse hierarchically structured data outlined earlier in section 2 of this review. The technique is designed to analyse longitudinal repeated measures data sets. One of its original applications was the investigation of educational performance in schools in which learning outcomes of pupils were charted (Aitkin *et al*., 1981; Raudenbush and Bryk, 1986; Nuttall *et al*., 1989). As pupils grow and develop, they move from class to class and learn more. Multilevel modelling can investigate the effects of influential factors on this learning such as teaching styles, age, gender and socio-economic background.

Multilevel modelling is an improvement on traditional analyses of repeated measures. In longitudinal data, a hierarchy is set at two levels: level 1 and level 2. Level 1 comprises the repeated measurement occasions whereas level 2 comprises the individual. Goldstein (1986) was probably

the first to apply the technique to a biological setting by exploring growth characteristics of children and adolescents. Since then, the software has been refined and is available commercially as MLwiN (Goldstein *et al*., 1998). In addition to describing the response of a group as a whole *i.e.* the mean response, MLwiN recognises and describes variation around the mean at both levels. In the context of growth for example, individuals have their own rates of growth that might vary randomly in comparison with the mean response of the group. Similarly, each individual's measurements might vary around their growth trajectory. Moreover, unlike traditional methods, data sets do not have to be complete from occasion to occasion and the time intervals between occasions do not have to be equal for each subject.

The analyses are comparatively new and complex but they provide a valuable adjunct to existing analytical techniques used in the physiology of exercise and kinanthropometry. After Goldstein's (1986) application to human biology, the first reported use in the physiology of exercise was by Baxter-Jones *et al*., (1993) who investigated cardiopulmonary function in elite child and adolescent athletes. However, within the last two years, there has been a marked increase in studies that have used MLwiN, especially where the interest is in the effects of children's growth and development on their responses and adaptations to exercise. Armstrong *et al*. (1999) used the technique to investigate longitudinal changes in peak oxygen uptake in 11-13 year old adolescents. Whereas Baxter-Jones *et al*. (1993) used an additive polynomial model, Armstrong *et al*. (1999) used a multiplicative allometric approach to overcome possible skewness and heteroscedasticity of error outlined earlier in this review. The results demonstrated age, sex and maturity effects that were independent of body mass on peak oxygen uptake. These effects were masked by the use of traditional ratio standards that simply divide peak oxygen uptake by body mass. Armstrong *et al*. (2000a) also successfully used multilevel modelling to investigate changes in maximal intensity exercise of young adolescents. There have been studies on physical activity patterns in 11-13 year-olds (Armstrong *et al*., 2000b), the effects of training on peak oxygen uptake and blood lipids in 13 to

14-year-old girls (Stoedefalke *et al*., 2000), maximal intensity exercise capability of 10 to 12 year-olds (De Ste Croix *et al*., 2001) and 12 to 17 year-olds (Armstrong *et al*., 2001), and peak oxygen uptake in relation to growth and maturation in 11 to 17 year-olds (Armstrong and Welsman, 2001). These studies have elegantly combined allometric and multilevel modelling which was especially relevant to the latter two studies where marked changes in body size were observed. A comparison of statistical techniques that can be used to interpret body size-related exercise performance during growth, including multilevel modelling has also been reported (Welsman and Armstrong, 2000).

Multilevel modelling has only recently been applied to sport and exercise science but it is probable that its use will increase. Published studies have demonstrated the utility of the technique, especially where the effects of growth, development, gender and training have to be disentangled. These and other confounding covariates influence for example, the effect of body mass on maximal and submaximal oxygen uptake and so impact on comparisons of physiological characteristics of groups that differ in size.

6. Summary

This review has highlighted biometric techniques that are especially relevant to the physiology of exercise and kinanthropometry and grouped them into four key areas. The section on bivariate correlation and linear and non-linear regression identified different regression models that can be used and ways in which appropriate models should be specified. The section on multiple regression highlighted the desirability of avoiding inclusion of variables that have no sound theoretical relationship with the criterion. Repeated measures designs feature prominently in physiological studies and can be complex. The section that included these types of design highlighted appropriate checks that should be made to ensure that underlying assumptions are not violated or, if they are, what corrective measures should be taken. Finally, examples of applications of multilevel modelling

to the physiology of exercise and kinanthropometry were presented. This is a comparatively recent analytical technique that is particularly suited to longitudinal studies that investigate physiological development from childhood, through adolescence to adulthood.

References

Aitkin, M., Anderson, D. and Hinde, J. (1981). Statistical modelling of data on teaching styles (with discussion). *Journal of the Royal Statistical Society A*, **144**, 148-161.

Armstrong, N. and Welsman, J.R. (2001). Peak $\dot{V}O_2$ in relation to growth and maturation in 11-17 year olds. *European Journal of Applied Physiology*, in press.

Armstrong, N., Welsman, J.R. and Chia, M.Y.H. (2001). Short term power output in relation to growth and maturation. *British Journal of Sports Medicine*, **35**, 118-124.

Armstrong, N., Welsman, J.R. and Kirby, B.J. (2000b). Longitudinal changes in 11-13-year-olds' physical activity. *Acta Pædiatrica*, **89**, 775-780.

Armstrong, N., Welsman, J.R., Nevill, A.M. and Kirby, B.J. (1999). Modeling growth and maturation changes in peak oxygen uptake in 11-13-year olds. *Journal of Applied Physiology*, **87**, 2230-2236.

Armstrong, N., Welsman, J.R., Williams, C.A. and Kirby, B.J. (2000a). Longitudinal changes in young people's short-term power output. *Medicine and Science in Sports and Exercise*, **32**, 1140-1145.

Atkinson, G. and Nevill, A.M. (2001) Selected issues in the design and analysis of sports performance research. *Journal of Sports Sciences*, this issue.

Bartlett, R. (1997). The use and abuse of statistics in sports and exercise sciences. *Journal of Sports Sciences*, **15**, 1-2.

Batterham, A.M. and George, K.P. (1997). Allometric modelling does not determine a dimensionless power function ratio for maximal muscular function. *Journal of Applied Physiology*, **83**, 2158-2166.

Baxter-Jones, A., Goldstein, H. and Helms, P. (1993). The development of aerobic power in young athletes. *Journal of Applied Physiology*, **75**, 1160-1167.

Biddle, S., Chatzisarantis, N., Gilbourne, D., Markland, D. and Sparkes, A. (2001) Research methods in sport psychology: quantitative and qualitative issues. *Journal of Sports Sciences*, this issue.

Bland, J.M. and Altman, D.G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*, **i**, 307-310.

Cockerill, I.M., Nevill, A.M. and Lyons, N. (1991). Modelling mood states in athletic performance. *Journal of Sports Sciences*, **9**, 205-212.

Cohen, J. and Cohen, P. (1983). *Applied Multiple Regression/Correlation Analysis for the Behavioural Sciences* (Second Edition), Hillsdale, NJ: Erlbaum.

De Ste Croix, M.B.A., Armstrong, N., Chia, M.Y.H., Welsman, J.R., Parsons, G. and Sharpe, P. (2001). Changes in short-term power output in 10- to 12-year-olds. *Journal of Sports Sciences*, **19**, 141-148.

Doyle, J.A. and Martinez, A.L. (1998). Reliability of a protocol for testing endurance performance in runners and cyclists. *Research Quarterly for Exercise and Sport*, **69**, 304-307.

El-Sayed, M.S., Balmer, J. Rattu, A.J.M. (1997). Carbohydrate ingestion improves endurance performance during a 1 h simulated cycling time trial. *Journal of Sports Sciences*, **15**, 223-230.

Eston, R.G., Cruz, A., Fu, F. and Fung, L. (1993). Fat-free mass estimation by bioelectrical impedance and anthropometric techniques in Chinese children. *Journal of Sports Sciences,* **11**, 241-247.

Eston, R.G., Fu, F. and Fung, L. (1995). Validity of conventional anthropometric techniques for estimating body composition in Chinese adults. *British Journal of Sports Medicine*, **29**, 52-56.

Eston, R.G. and Peters, D. (1999). Effects of cold water immersion on the symptoms of exercise induced muscle damage. *Journal of Sports Sciences*, **17**, 231-238.

Eston, R.G., Robson, S. and Winter, E.M. (1993). A comparison of oxygen uptake during running in children and adults. In *Kinanthropometry IV* (edited by J.W. Duquet and J.A.P. Day), London: E & FN Spon *pp*. 236-241.

Eston, R.G., Rowlands, A.V. and Ingledew, D.K. (1998). Validity of heart rate, pedometry and accelerometry for predicting the energy cost of children's activities. *Journal of Applied Physiology,* **84**, 362-371

Eston, R.G., Shephard, S., Kreitzman, S., Coxon, A., Brodie, D.A., Lamb, K.L., and Baltzopoulos, V. (1992). Effect of very low calorie diet on body composition and exercise response in sedentary women. *European Journal of Applied Physiology*, **65**, 452-458.

Eston, R.G., Winter, E.M. and Baltzopoulos, V. (1997). Ratio standards and allometric modelling to scale peak power output for differences in lean upper leg volume in men and women. *Journal of Sports Sciences*, **15**, 29.

Franks, B.D. and Huck, S.W. (1986). Why does everyone use the .05 significance level? *Research Quarterly for Exercise and Sport*, **57**, 245-249.

Gill, J.M.R., Murphy, M.H. and Hardman, A.E. (1998). Postprandial lipemia: effects of intermittent versus continuous exercise. *Medicine and Science in Sports and Exercise*, **30**, 1515-1520.

Goldstein, H. (1986). Efficient statistical modelling of longitudinal data. *Annals of Human Biology*, **13**, 129-141.

Goldstein, H., Rasbash, J., Plewis, I., Draper, D., Browne, W., Yang, M., Woodhouse, G. and Healy, M. (1998). *A User's Guide to MLwiN*. London: Institute of Education, University of London.

Houtkooper L.B., Lohman T.G., Going S.B. and Hall M.C. (1989). Validity of bioelectrical impedance for body composition assessment in children. *Journal of Applied Physiology*, **66**, 814-821

Howell, D.C. (1997). *Statistical Methods for Psychology*. Belmont: Duxbury.

Huck, S.W., Cormier, W.H. and Bounds, W.G. (1974). *Reading Statistics and Research.* New York: Harper and Row.

Huck, S.W. and Cormier, W.H. (1996). *Reading Statistics and Research*. New York: Harper Collins.

Huxley, J.S. (1932). *Problems of Relative Growth*. New York: Dial.

Jackson, A.S and Pollock, M.L. (1978). Generalized equations for predicting body density of men. *British Journal of Nutrition*, **40**, 497-504.

Katch, V. (1973). Use of oxygen/body weight ratio in correlation analysis: spurious correlations and statistical considerations. *Medicine and Science in Sports*, **5**, 223-257.

Kermack, K.A. and Haldane, J.B.S. (1950). Organic correlation and allometry. *Biometrika*, **37**, 30-41.

Kinnear, P.R. and Gray, C.D. (1997). *SPSS for Windows Made Simple* (Second Edition). Hove: Psychology Press.

Kuhry, B. and Marcus, L.F. (1977). Bivariate linear models in biometry. *Systematic Zoology*, **26**, 201-209.

Lamb, K.L., Eston, R.G. and Corns, D. (1999). The reliability of ratings of perceived exertion during progressive treadmill exercise. *British Journal of Sports Medicine*, **33**, 336-339.

Louie, L., Eston, R.G., Rowlands, A., Tong, T., Ingledew, D.K. and Fu, F. (1999). Validity of heart rate, pedometry and accelerometry for predicting the energy expenditure in Hong Kong Chinese children. *Pediatric Exercise Science*, **11**, 229-239.

Morris, J.G., Nevill, M.E., Lakomy, H.K.A., Nicholas, C. and Williams, C. (1998). Effect of a hot environment on performance of prolonged, intermittent, high-intensity shuttle running. *Journal of Sports Sciences*, **16**, 677-686.

Mullineaux, D.R., Bartlett, R.M. and Bennett, S. (2001) Research design and statistics in biomechanics and motor control. *Journal of Sports Sciences*, this issue.

Munro, B.H. (2001). *Statistical Methods for Health Care Research* (Fourth Edition). Philadelphia: Lippincott Williams and Wilkins.

Nevill, A.M. (1994). The need to scale for differences in body size and mass: an explanation of Kleiber's 0.75 mass exponent. *Journal of Applied Physiology*, **77**, 2870-2873.

Nevill, A.M. and Atkinson, G. (2001). Statistical methods in kinanthropometry and exercise physiology. In *Kinanthropometry and Exercise Physiology Laboratory Manual: Tests, Procedures and Data* (Second Edition). (Edited by R.G. Eston and T. Reilly). London: Routledge, in press.

Nevill, A.M. and Holder, R.L. (1994). Modelling maximum oxygen uptake: a case study in non-linear regression formulation and comparison. *Journal of the Royal Statistical Society, Series C*, **43**, 653-666.

Nevill, A.M. and Holder, R.L. (1995). Scaling, normalizing and per ratio standards: an allometric modelling approach. *Journal of Applied Physiology*, **79**, 1027-1031.

Nevill, A.M. and Holder, R.L. (1999). Identifying population differences in lung function: results from the Allied Dunbar national fitness survey. *Annals of Human Biology*, **26**, 267-285.

Norris, S.R. and Petersen, S.R. (1998). Effects of endurance training on transient oxygen uptake responses in cyclists. *Journal of Sports Sciences*, **16**, 733-738.

Nuttall, D., Goldstein, H., Prosser, R. and Rasbash, J. (1989). Differential school effectiveness. *International Journal of Educational Research*, **13**, 769-776.

Owens, S., Gutin, B., Allison, J., Riggs, S., Ferguson, M., Litaker, M. and Thompson, W. (1999). Effect of physical training on total and visceral fat in obese children. *Medicine and Science in Sports and Exercise,* **31**, 143-148.

Packard, G.J. and Boardman, T.J. (1987). The misuse of ratios to scale physiological data that vary allometrically with body size. In *New Directions in Ecological Physiology* (edited by M.E. Feder, A.F. Bennett, W.W. Burggren and R.B. Huey) Cambridge: Cambridge University Press, *pp*. 216-239.

Perneger, T.V. (1998). What is wrong with Bonferroni adjustments? *British Medical Journal*, **316**, 1236-1238.

Peyrebrune, M.C., Nevill, M.E., Donaldson, F.J. and Cosford, D.J. (1998). The effects of oral creatine supplementation on performance in single and repeated sprint swimming. *Journal of Sports Sciences*, **16**, 271-279.

Popper, K.R. (1963). *Conjectures and Refutations*. London: Routledge and Kegan Paul.

Raudenbush, S. and Bryk, A. (1986). A hierarchical model for studying school effects. *Sociology of Education*, **59**, 1-17.

Rayner, J.M.V. (1985). Linear relations in biomechanics: the statistics of scaling functions. *Journal of Zoology (London) Series A*, **206**, 415-439.

Ricker, W.E. (1973). Linear regression in fishery research. *Journal of the Fisheries Research Board Canada*, **30**, 409-434.

Rogers, D.M., Olson, B.L. and Wilmore, J.H. (1995). Scaling for the $\dot{V}O_2$-to-body size relationship among children and adults. *Journal of Applied Physiology*, **79**, 958-967.

Rowlands, A.V., Eston, R.G., and Ingledew, D.K. (1999). The relationship between activity levels, aerobic fitness, and body fat in 8-10 year old children. *Journal of Applied Physiology*, **86**, 1428-1435.

Sale, D.G. (1991). Testing strength and power. In *Physiological Testing of the High-Performance Athlete* (Second Edition), (edited by J.D. MacDougall, H.A. Wenger and H.J. Green), Champaign IL: Human Kinetics *pp*. 21-106.

Schmidt-Nielsen, K. (1984). *Scaling: Why is Animal Size so Important?* Cambridge: Cambridge University Press.

Seim, E. and Sæther, B-E. (1983). On rethinking allometry: which regression model to use? *Journal of Theoretical Biology*, **104**, 161-168.

Siegel, S. and Castellan, N.J. (1988). *Nonparametric Statistics for the Behavioural Sciences* (Second Edition), New York: McGraw-Hill.

Smith, R.J. (1984). Allometric scaling in comparative biology: problems of concept and method. *American Journal of Physiology*, **246**, R152-R160.

Snedecor, G.W. and Cochran, W.G. (1989). *Statistical Methods* (Eighth Edition). Ames, IO: Iowa State University Press.

Sokal, R.R. and Rohlf, F.J. (1995). *Biometry* (Third Edition). New York: W.H. Freeman and Company.

Stevens, J. (1996). *Applied Multivariate Statistics for the Social Sciences*. Mahwah, NJ: Lawrence Erlbaum Associates.

Stoedefalke, K., Armstrong, N., Kirby, B.J. and Welsman, J.R. (2000). Effect of training on peak oxygen uptake and blood lipids in 13 to 14-year-old girls. *Acta Pædiatrica*, **89**, 1290-1294.

Tabachnick, B.G. and Fidell, L.S. (1996). *Using Multivariate Statistics* (Third Edition). New York: Harper Collins.

Tanner, J.M. (1949). Fallacy of per-weight and per-surface area standards and their relationship with spurious correlation. *Journal of Applied Physiology*, **2**, 1-15.

Teissier, G. (1948). La relation d'allométrie: sa significance statistique et biologique. *Biometrics*, **4**, 14-48.

Thomas, D.O., Bowdoin, B.A., Brown, D.D., and McCaw, S.T. (1998). Nasal strips and mouthpieces do not affect power output during anaerobic exercise. *Research Quarterly for Exercise and Sport*, **69**, 201-204

Thomas, J.R. and Nelson, J.K. (1996). *Research Methods in Physical Activity* (Third Edition). Champaign, IL: Human Kinetics.

Thompson, D'A. W. (1917). *Growth and Form*. Cambridge: Cambridge University Press.

Trappe, T.A., Starling, R.D., Jozsi, A.C., Goodpaster, B.H., Trappe, S.W., Nomura, T., Obaru, S. and Costill, D.L. (1995). Thermal responses to swimming in three water temperatures: influence of a wet suit. *Medicine and Science in Sports and Exercise*, **27**, 1014-1021.

Vanderburgh, P.M., Mahar, M.T. and Chou, C.H. (1995). Allometric scaling of grip strength by body mass in college-age men and women. *Research Quarterly for Exercise and Sport*, **66**, 80-84.

Vincent, W.J. (1999). *Statistics in Kinesiology* (Second Edition). Champaign, IL: Human Kinetics.

Welsman, J.R. and Armstrong, N. (2000). Statistical techniques for interpreting body size-related exercise performance during growth. *Pediatric Exercise Science*, **12**, 112-127.

Welsman, J.R., Armstrong, N., Kirby, B.J., Nevill, A.M. and Winter, E.M. (1996). Scaling peak $\dot{V}O_2$ for differences in body size. *Medicine and Science in Sports and Exercise*, **28**, 259-265.

Whipp, B.J. (1996). Domains of aerobic function and their limiting parameters. In: *The Physiology and Pathophysiology of Exercise Tolerance* (edited by J.M. Steinacker and S.A. Ward), pp 83-89. New York: Plenum Press.

Winter, E.M. (1996). Importance and principles of scaling for size differences. In *The Child and Adolescent Athlete* (edited by O. Bar-Or), Oxford: Blackwell, *pp*. 673-679.

Winter, E.M., Brookes, F.B.C. and Hamley, E.J. (1991). Maximal exercise performance and lean leg volume in men and women. *Journal of Sports Sciences*, **9**, 3-13.

Winter, E.M., Brown, D., Roberts, N.K.A., Brookes, F.B.C. and Swaine, I.L. (1996). Optimized and corrected peak power output during friction-braked cycle ergometry. *Journal of Sports Sciences*, **14**, 513-521.

Winter, E.M. and Nevill, A.M. (1996). Scaling: adjusting for differences in body size. In: *Kinanthropometry and Exercise Physiology Laboratory Manual* (edited by R. Eston and T. Reilly), London: E & FN Spon, *pp*. 321-335.

Table 1. Mauchly's Test of Sphericity

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon(a)		
					Greenhouse-Geisser	Huynh-Feldt	Lower-bound
TRIALS	1.000	.000	0	.	1.000	1.000	1.000
LEVELS	.183	23.273	5	.000	.505	.549	.333
TRIALS * LEVELS	.385	13.085	5	.023	.629	.717	.333

Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.

a May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the layers (by default) of the Tests of Within Subjects Effects table.

Table 2. Tests of Within-Subjects Effects

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
TRIALS	Sphericity Assumed	1.531	1	1.531	.586	.456
	Greenhouse-Geisser	1.531	1.000	1.531	.586	.456
	Huynh-Feldt	1.531	1.000	1.531	.586	.456
	Lower-bound	1.531	1.000	1.531	.586	.456
LEVELS	Sphericity Assumed	504.625	3	168.208	358.314	.000
	Greenhouse-Geisser	504.625	1.514	333.313	358.314	.000
	Huynh-Feldt	504.625	1.648	306.225	358.314	.000
	Lower-bound	504.625	1.000	504.625	358.314	.000
TRIALS * LEVELS	Sphericity Assumed	5.344	3	1.781	5.764	.002
	Greenhouse-Geisser	5.344	1.888	2.831	5.764	.009
	Huynh-Feldt	5.344	2.151	2.484	5.764	.006
	Lower-bound	5.344	1.000	5.344	5.764	.030

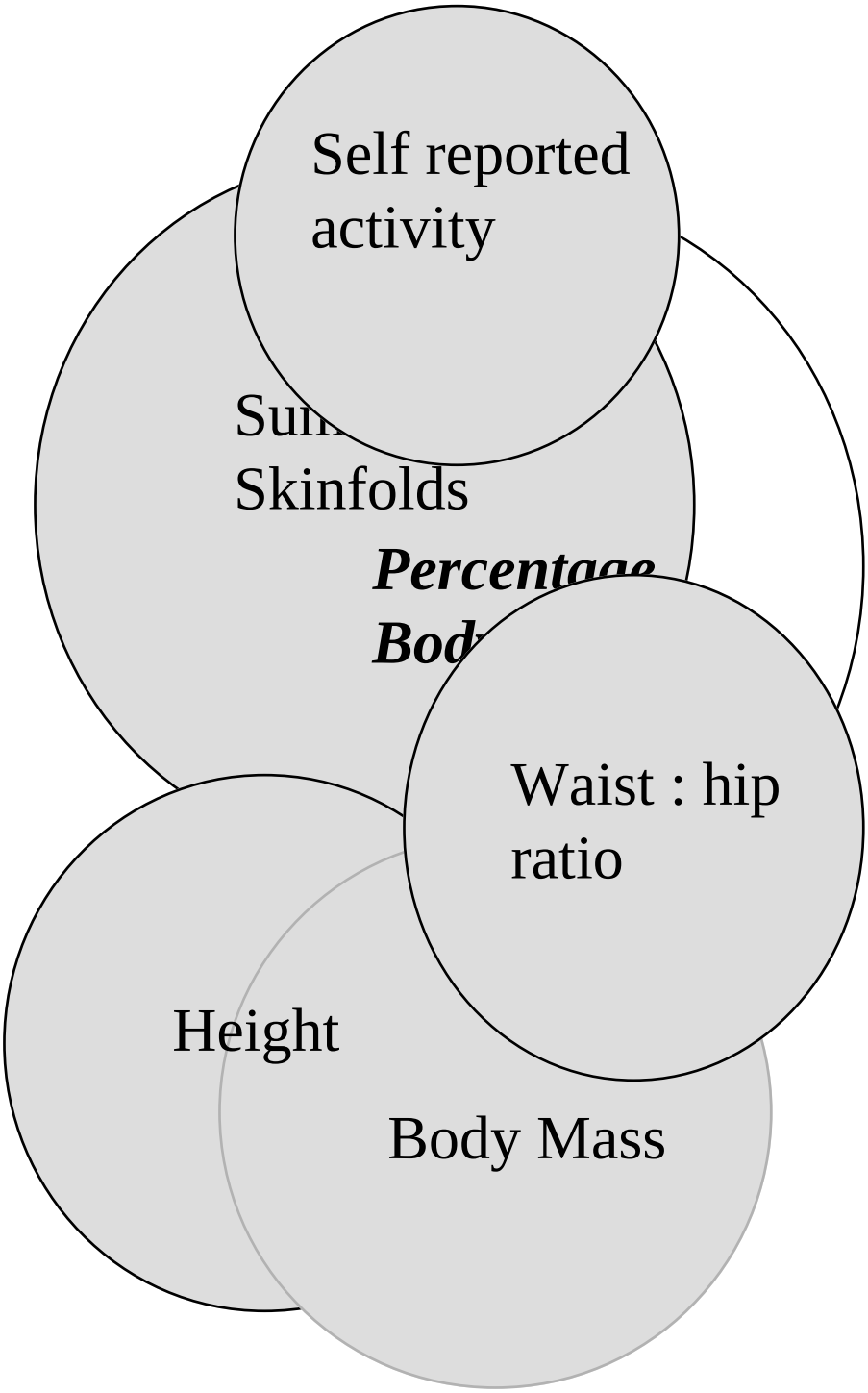
Self reported activity
Sum
Skinfolds
Percentage
Bod
Waist : hip ratio
Height
Body Mass

Figure Captions

Figure 1. Regression lines of *Y* on *X* (line A), *X* on *Y* (line B) and geometric mean (line C). For explanation, see text.

Figure 2. Venn diagram to exemplify how an independent variable is selected in multiple regression analysis based on the relationship with the criterion and other independent variables.